Compression of Generative Pre-trained Language Models via Quantization

Year: 2022 | Citations: 112 | Authors: Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang

---

Abstract

The increasing size of generative Pre-trained Language Models (PLMs) have greatly increased the demand for model compression. Despite various methods to compress BERT or its variants, there are few attempts to compress generative PLMs, and the underlying difficulty remains unclear. In this paper, we compress generative PLMs by quantization. We find that previous quantization methods fail on generative tasks due to the homogeneous word embeddings caused by reduced capacity and the varied distribution of weights. Correspondingly, we propose a token-level contrastive distillation to learn distinguishable word embeddings, and a module-wise dynamic scaling to make quantizers adaptive to different modules. Empirical results on various tasks show that our proposed method outperforms the state-of-the-art compression methods on generative PLMs by a clear margin. With comparable performance with the full-precision models, we achieve 14.4x and 13.4x compression rate on GPT-2 and BART, respectively.